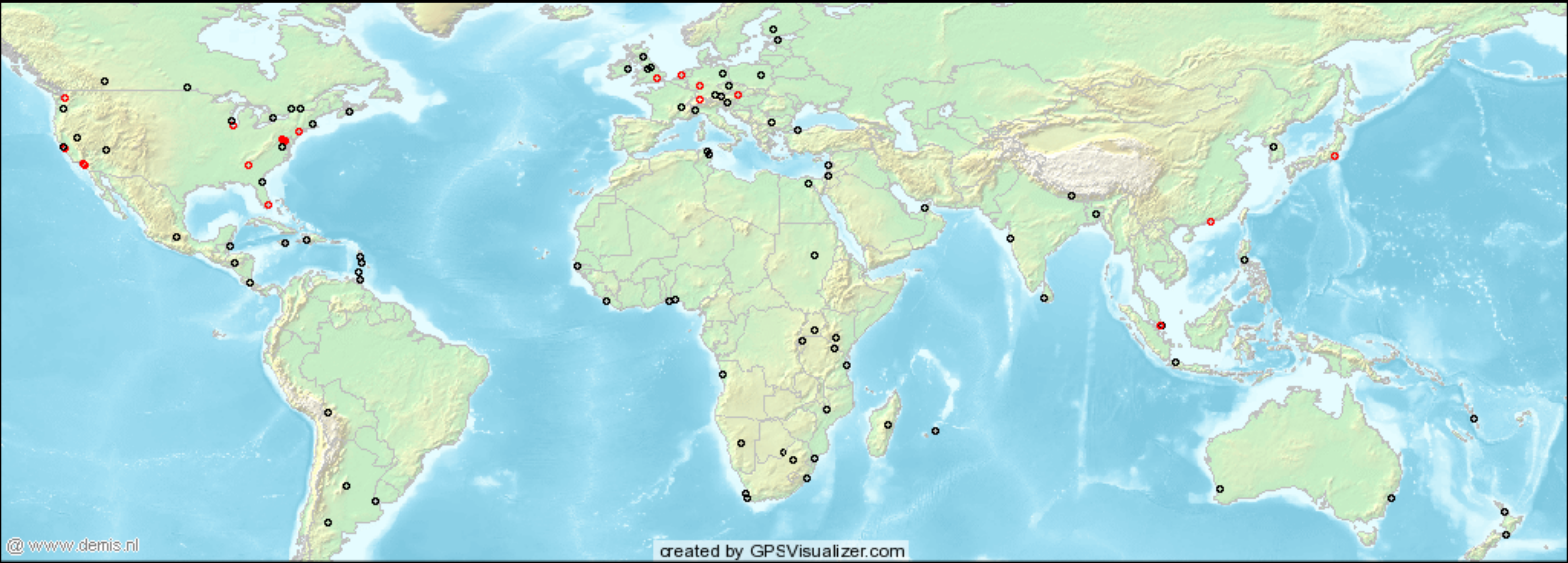
Tier-1's break Anycast DNS

Zhihao Li, Neil Spring

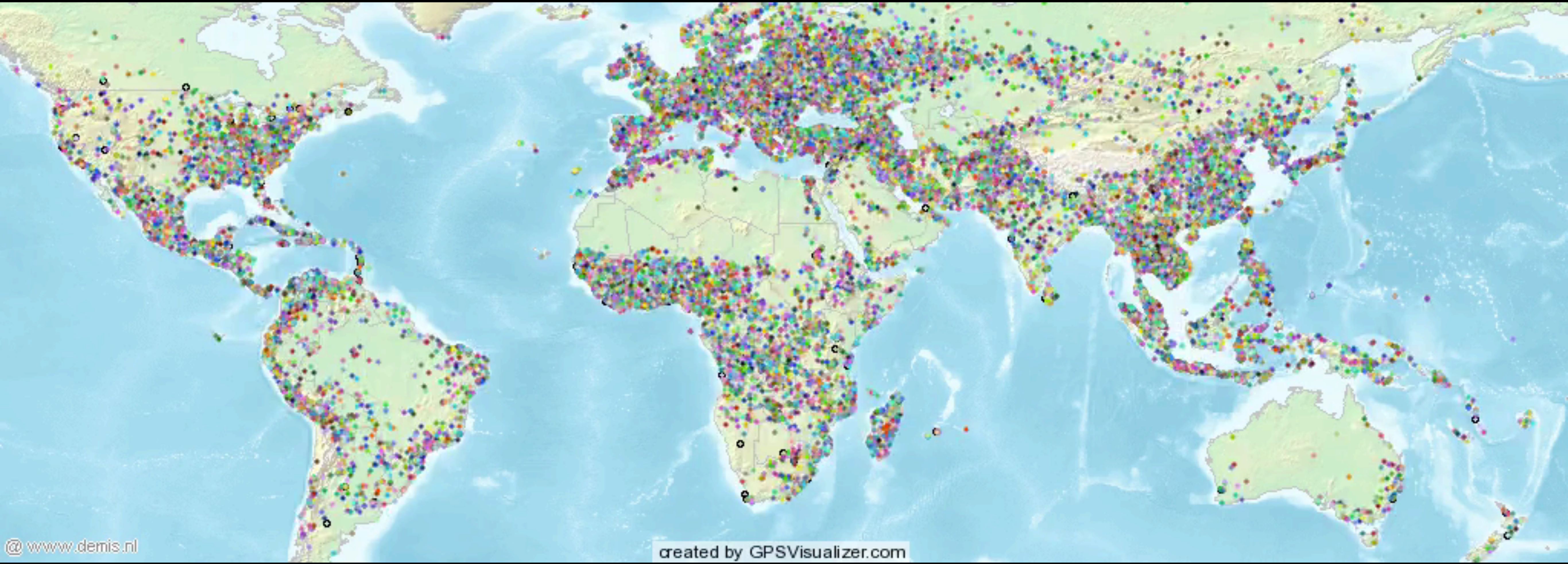
D-Root: 199.7.91.13

- 111 Anycast replicas:
  - 19 global (red): advertised without restriction
  - 92 local (black): advertised one hop in BGP



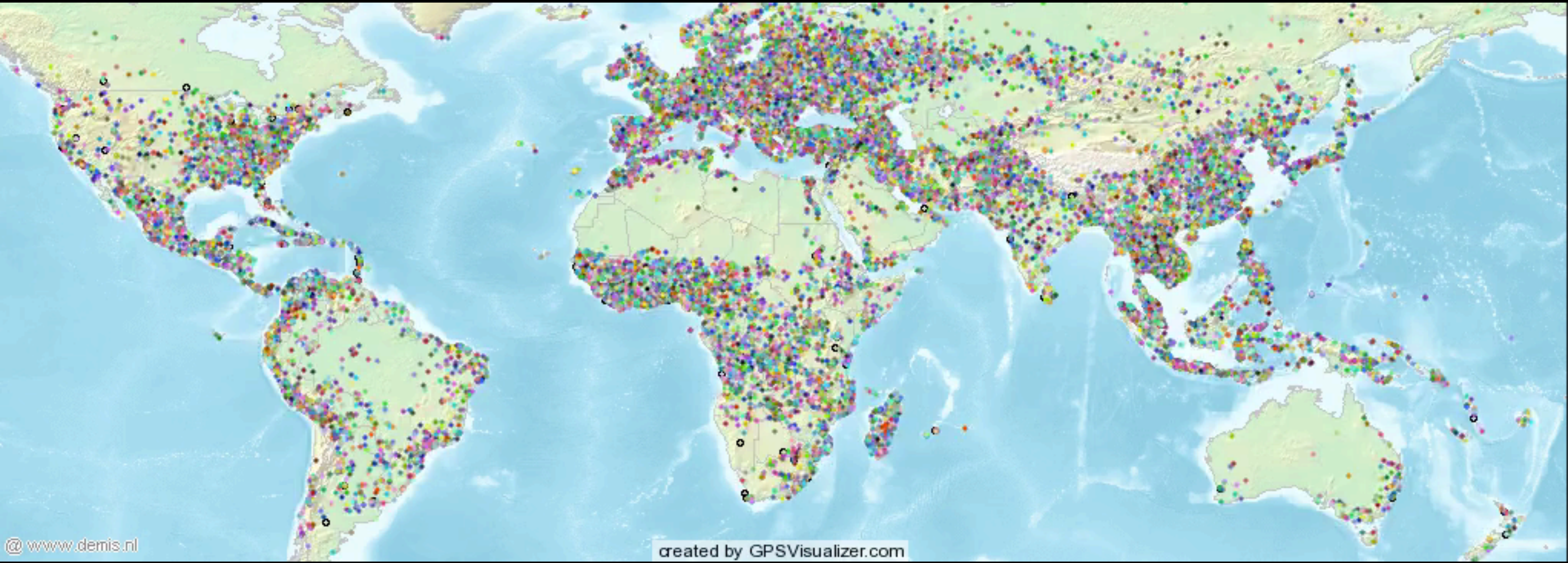
Anycast

- Mental model:
  - Packets sent to an anycast address travel to the nearest* replica, subject to global/local constraints.
  - More replicas should mean lower latency, better distribution, reliability against denial-of-service attacks.



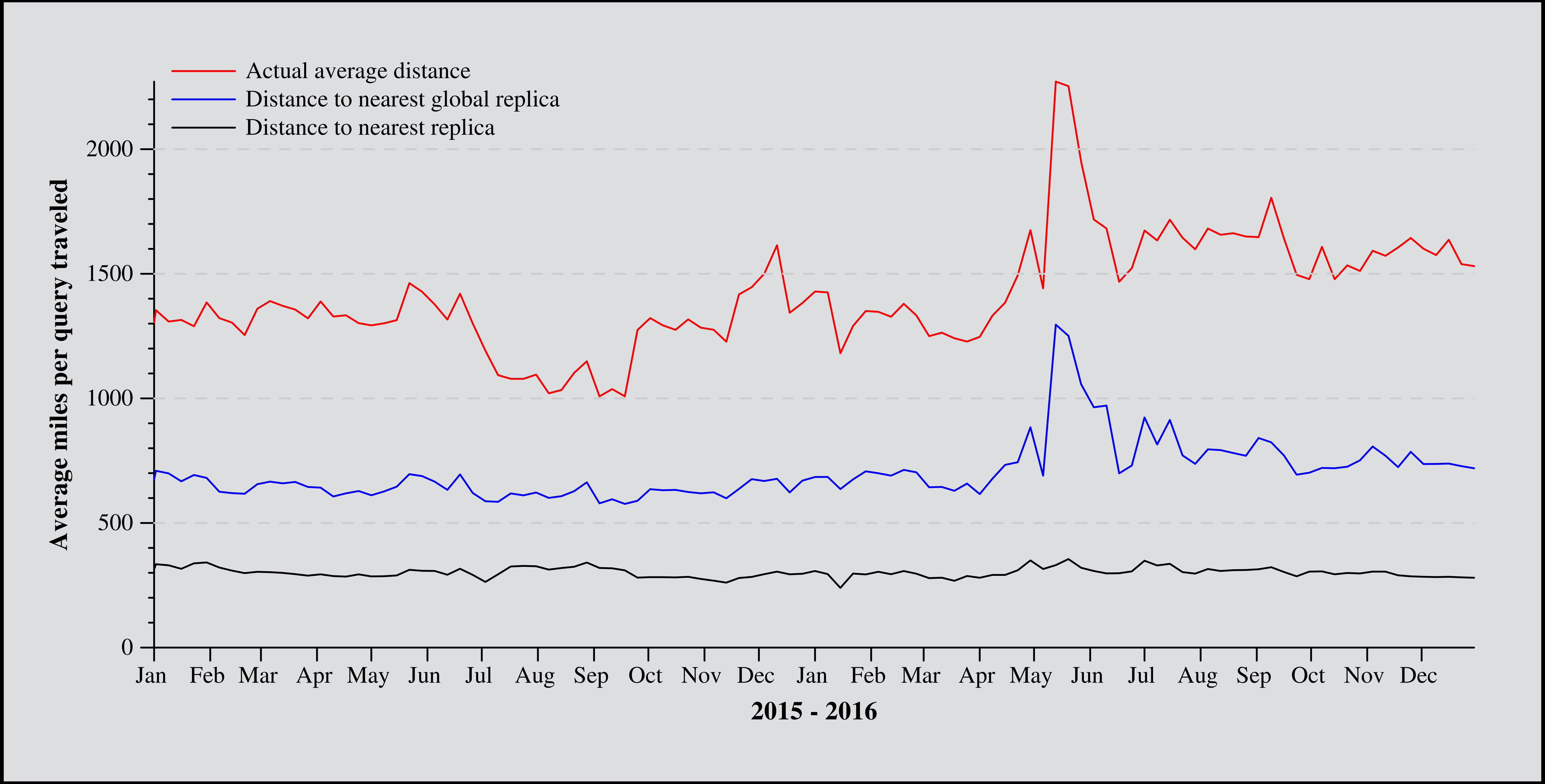
Anycast

- Mental model:
  - Packets sent to an anycast address travel to the nearest* replica, subject to global/local constraints.
  - More replicas should mean lower latency, better distribution, reliability against denial-of-service attacks.



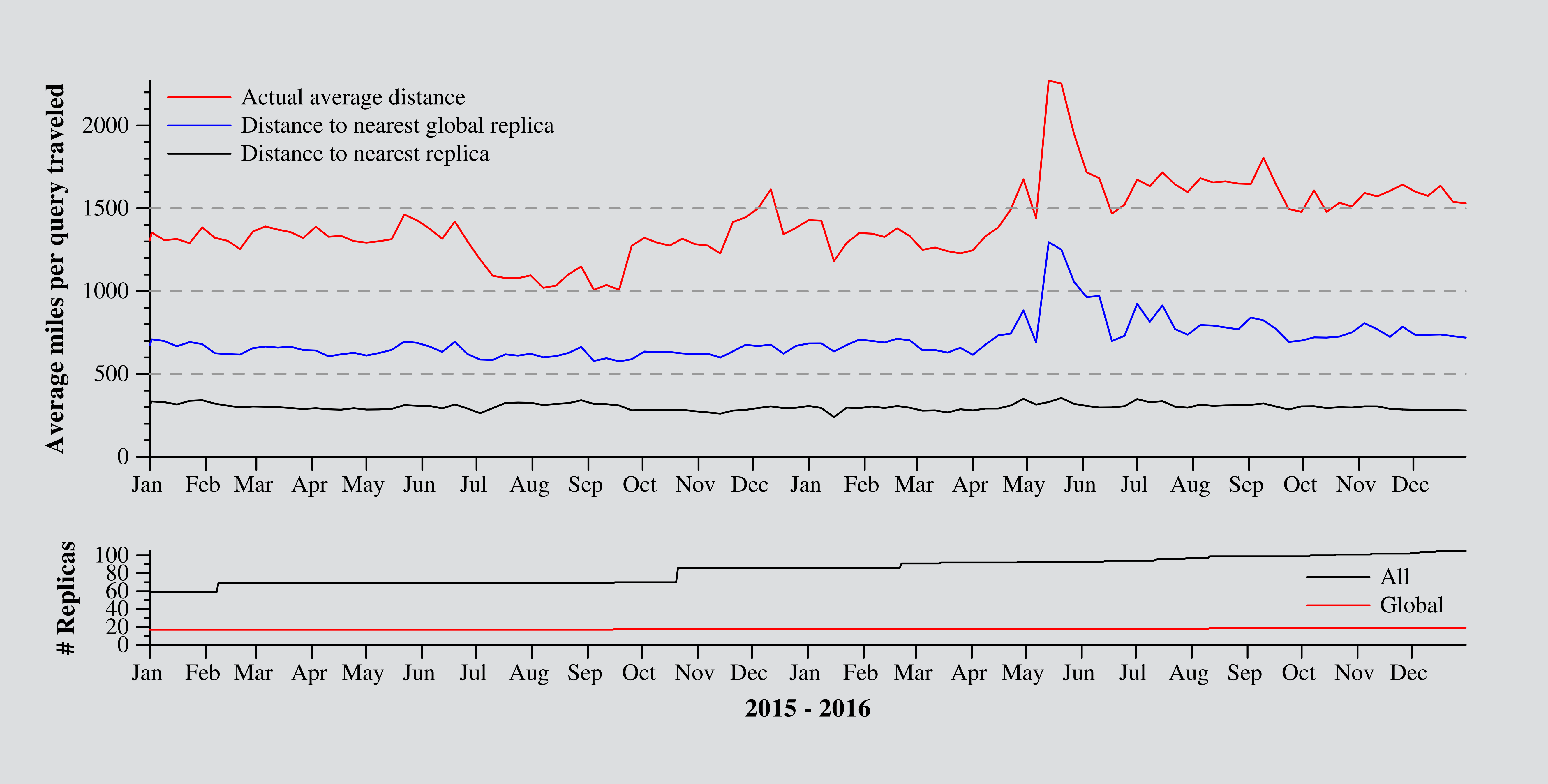
Reality

- 4-5x optimal delay (to a local), 2x expected (nearest global)



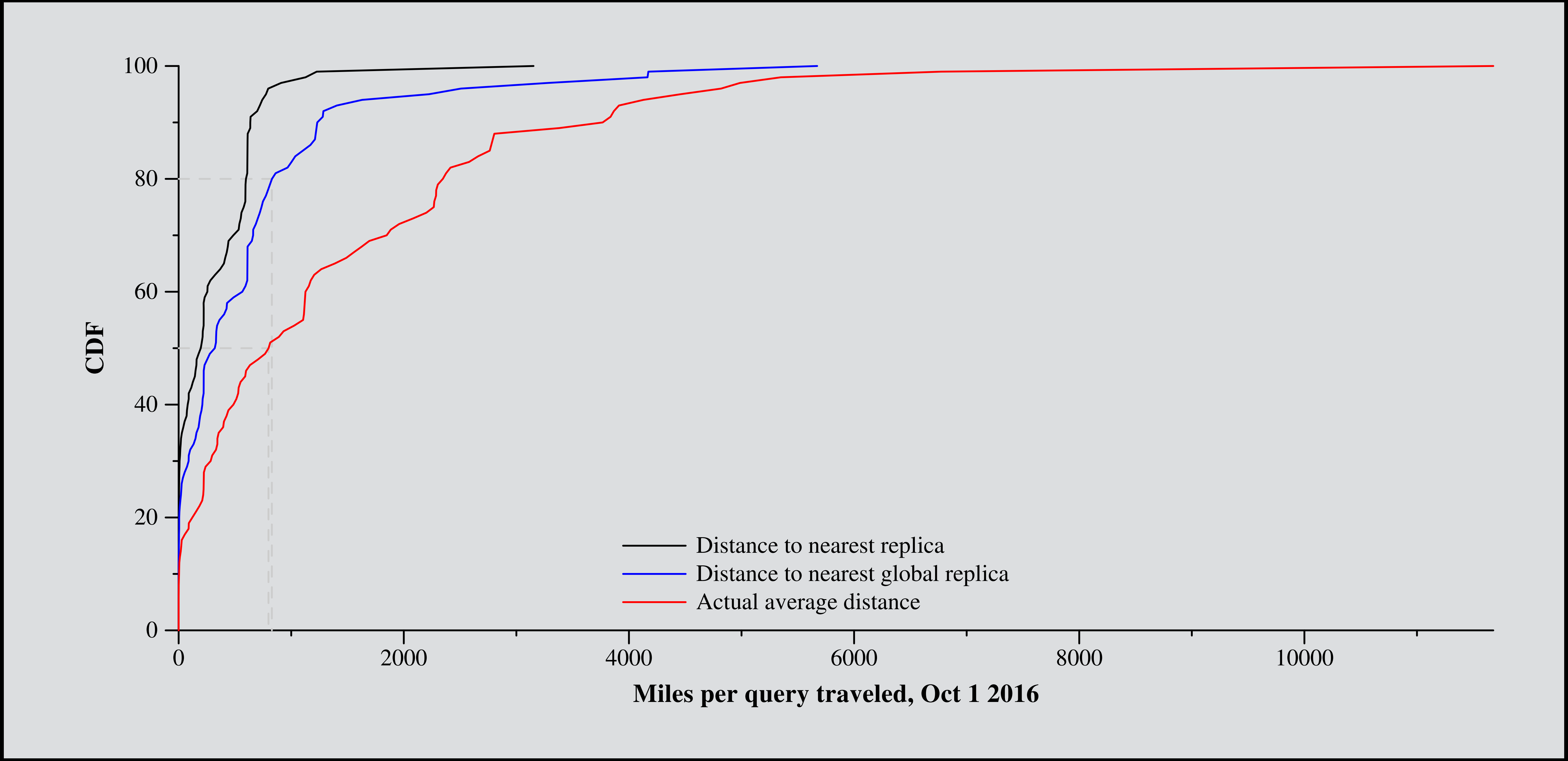
Reality

- Despite doubling the number of (local) replicas



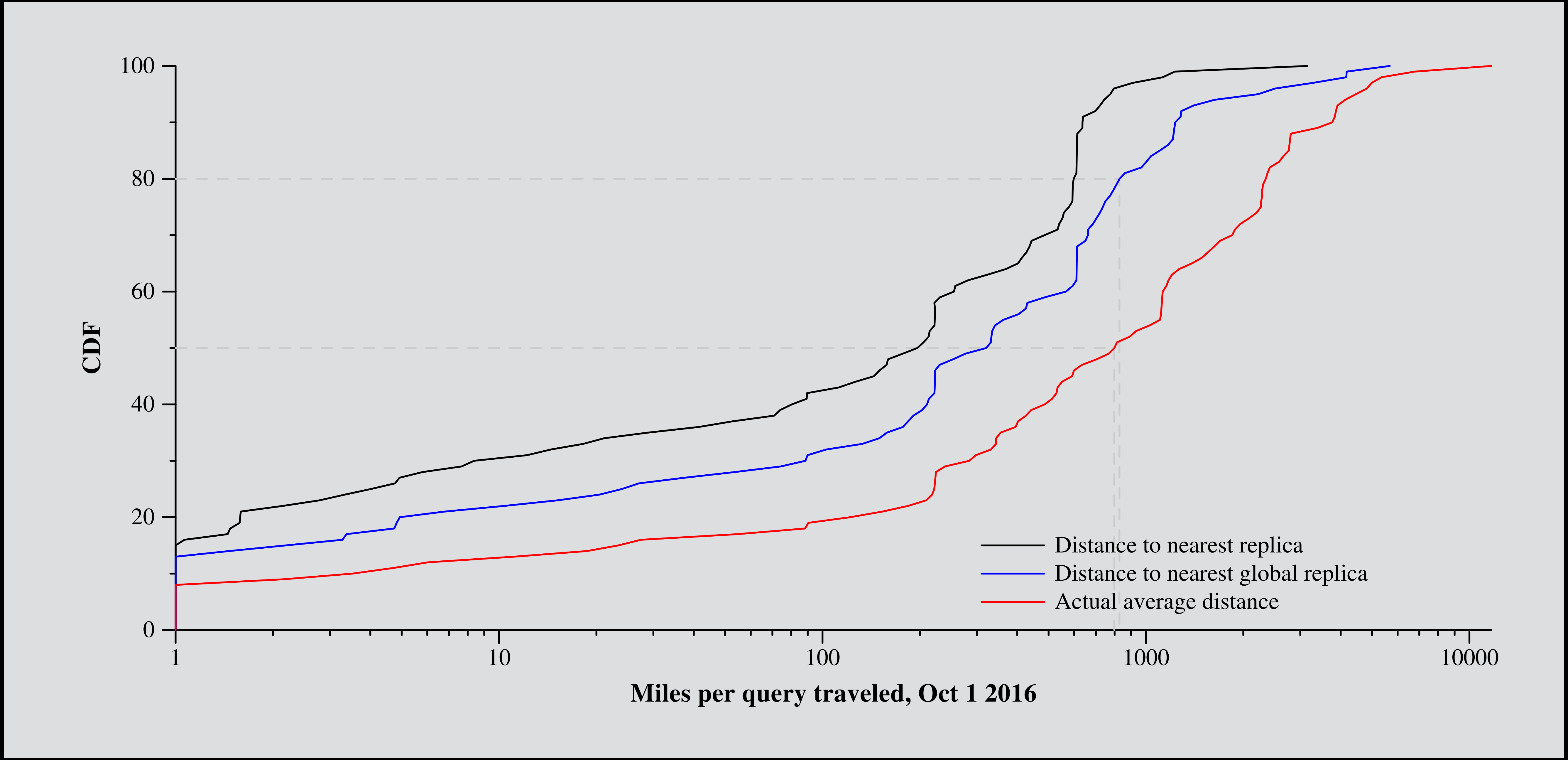
Reality

- 80% of queries should take under 1000 miles (16ms RTT)
- 50% are traveling farther.



Reality

- Same data, first week in Oct 2016, log scale x-axis.
- Even when there's a global replica in your city…



How do we fix it?

- More sites?
- More peerings?
- Better policies?
- Make local replicas global?

- What if ISPs chose cleverly from their providers?
  - Pathological behavior must be atypical, right?

- Is it even broken?

Similar observations

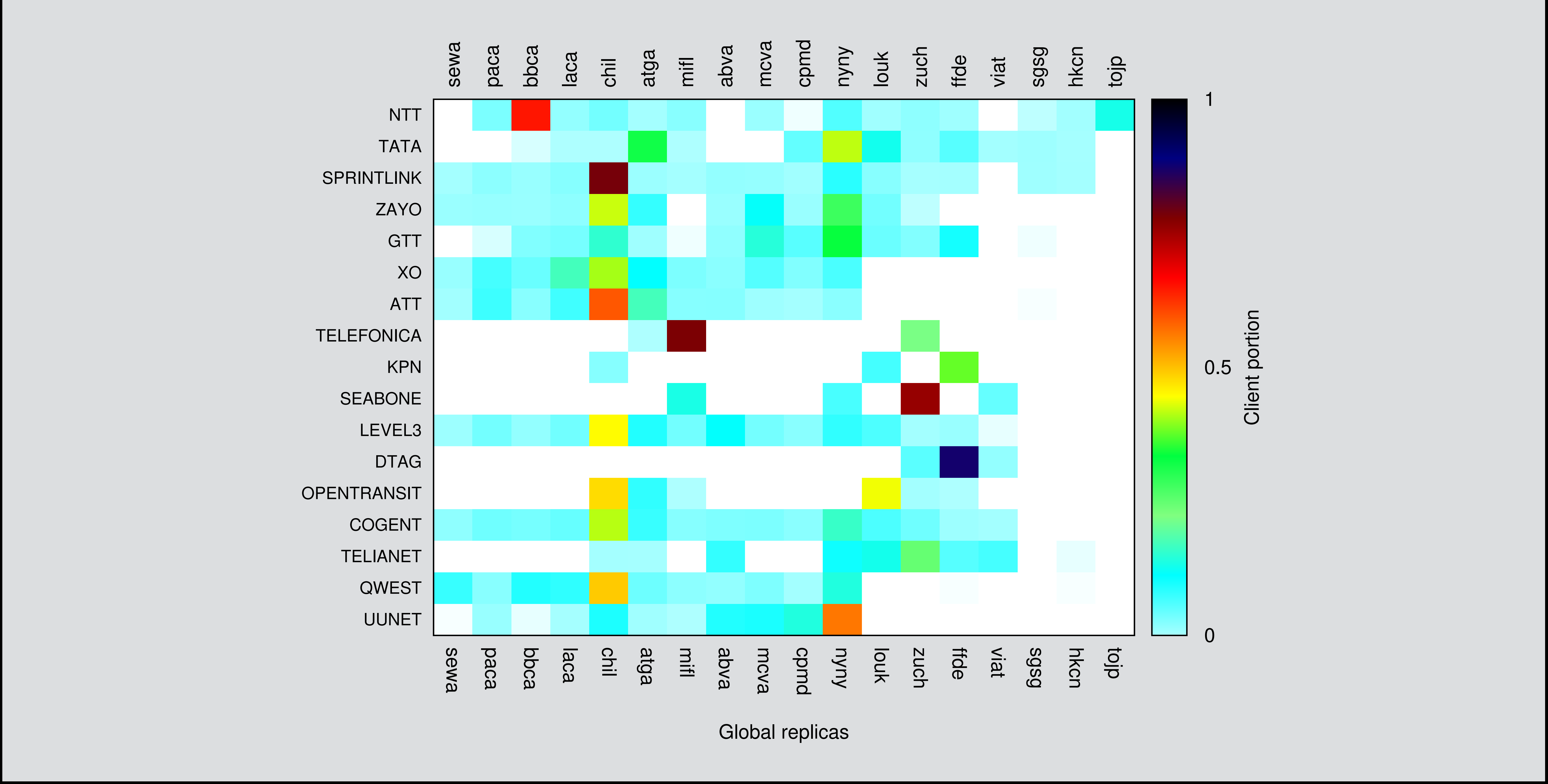
- *Anycast Latency: How Many Sites Are Enough?* Schmidt, Heidemann, Kuipers
  - Used Atlas probes (not traces) to look at C, F, K, L root.
  - More sites doesn’t correlate with lower latency
  - Making local sites global didn’t help K

It’s the tier-1’s

(I think)

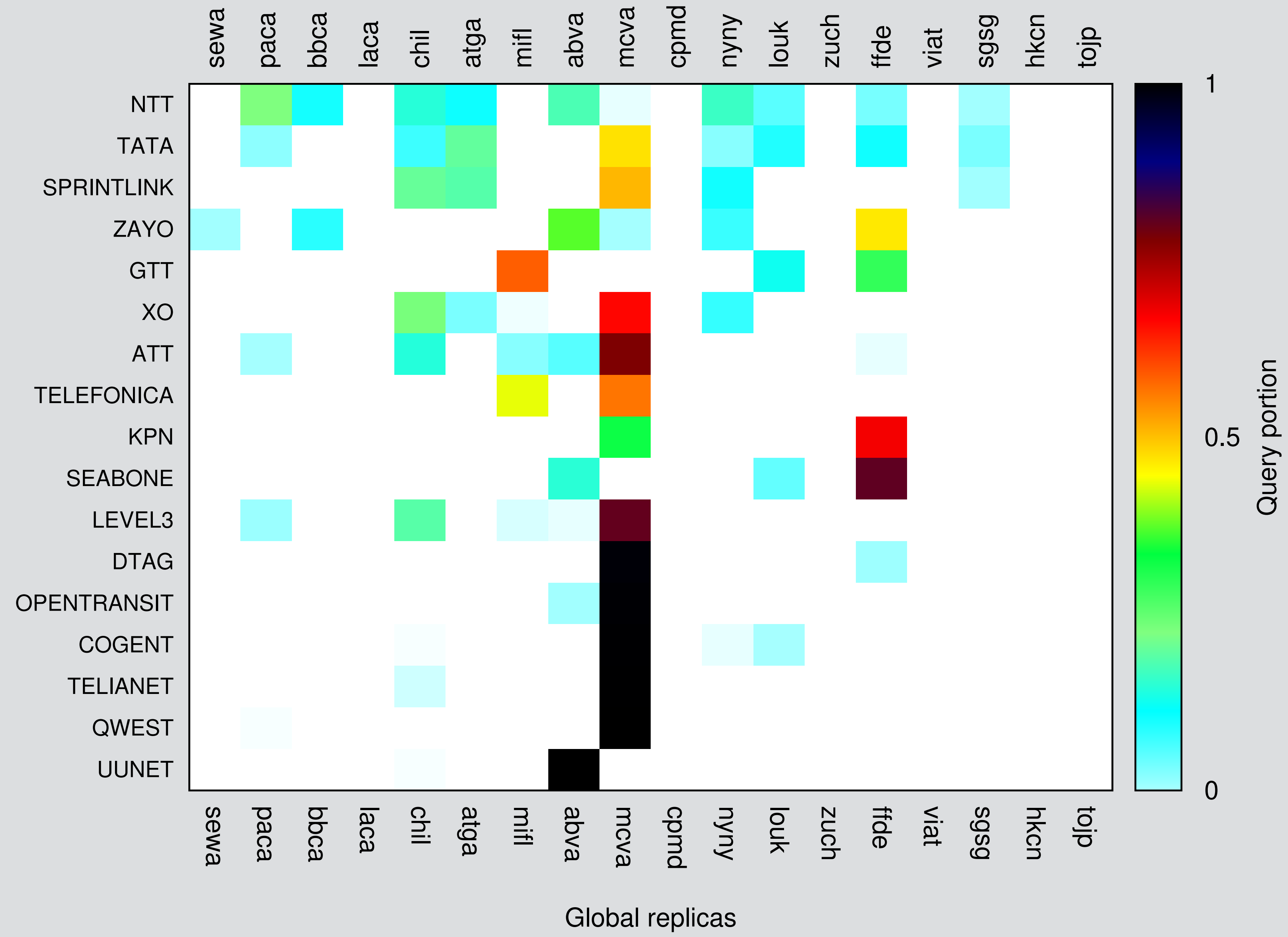
Source (resolver) location

- For addresses originated by Tier 1’s, what is their nearest replica. Intensity by query volume.



Request destination

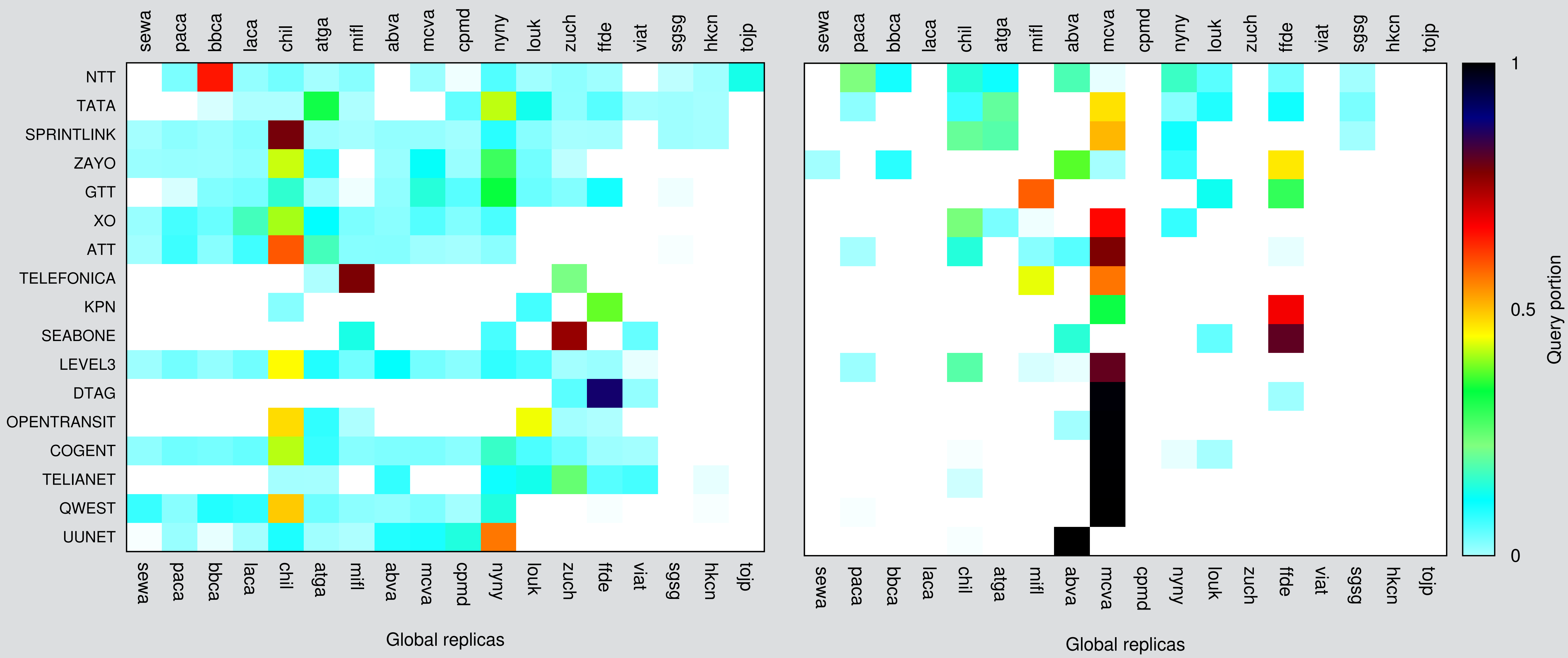
- For addresses originated by Tier 1's, what is their chosen replica. Intensity by query volume.



Would you like to see them again?

Often McLean, VA.

- Traffic from tier-1 address space *can* arrive on other replicas, but generally does not.



Could just be us.

Could just be us.

No.

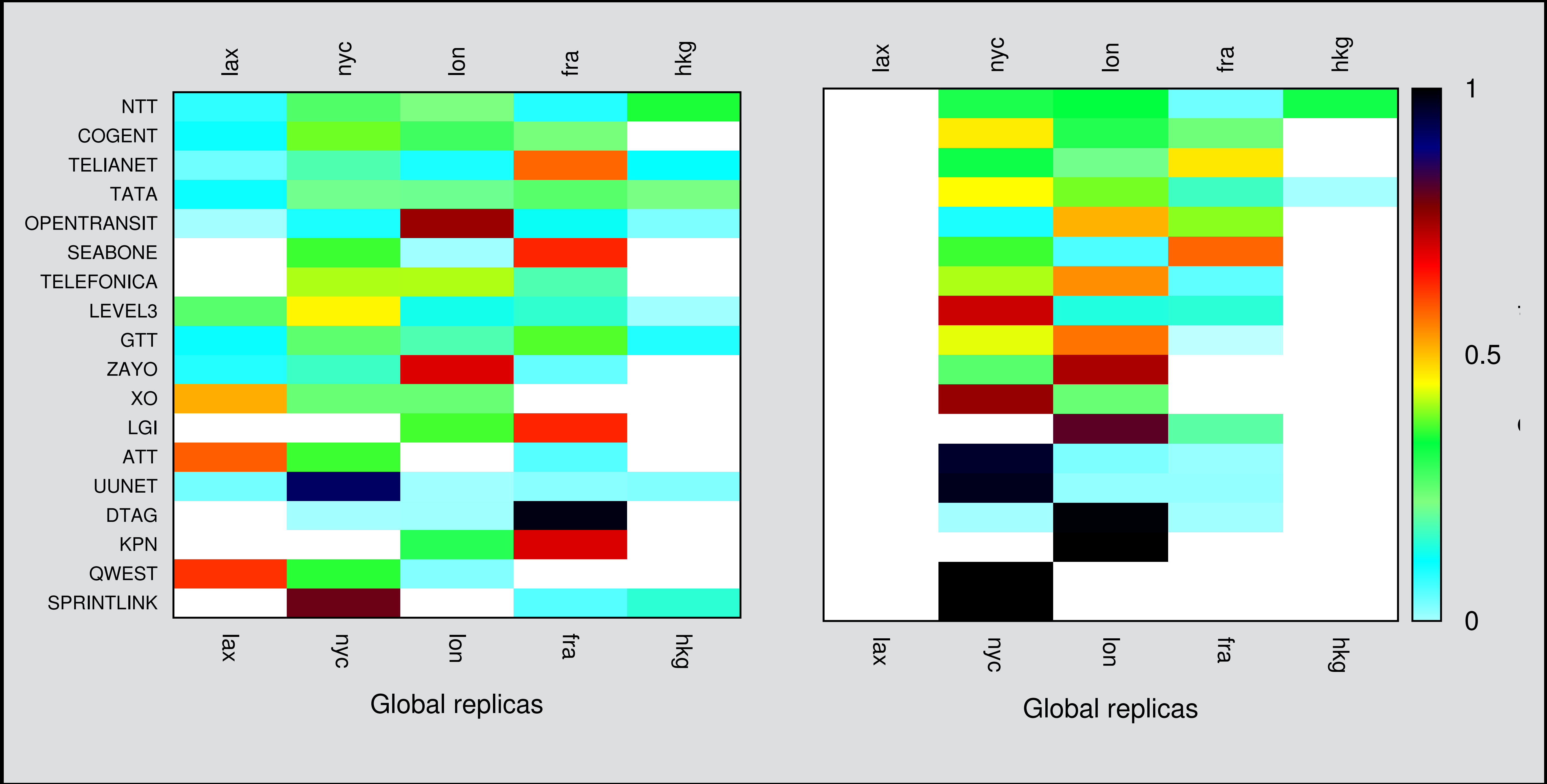
Could just be us.

No.

This time using RIPE Atlas data, same Oct 1, 2016.
Now counting vantage points whose queries transit a tier-1
(since we have traceroutes) instead of queries received.

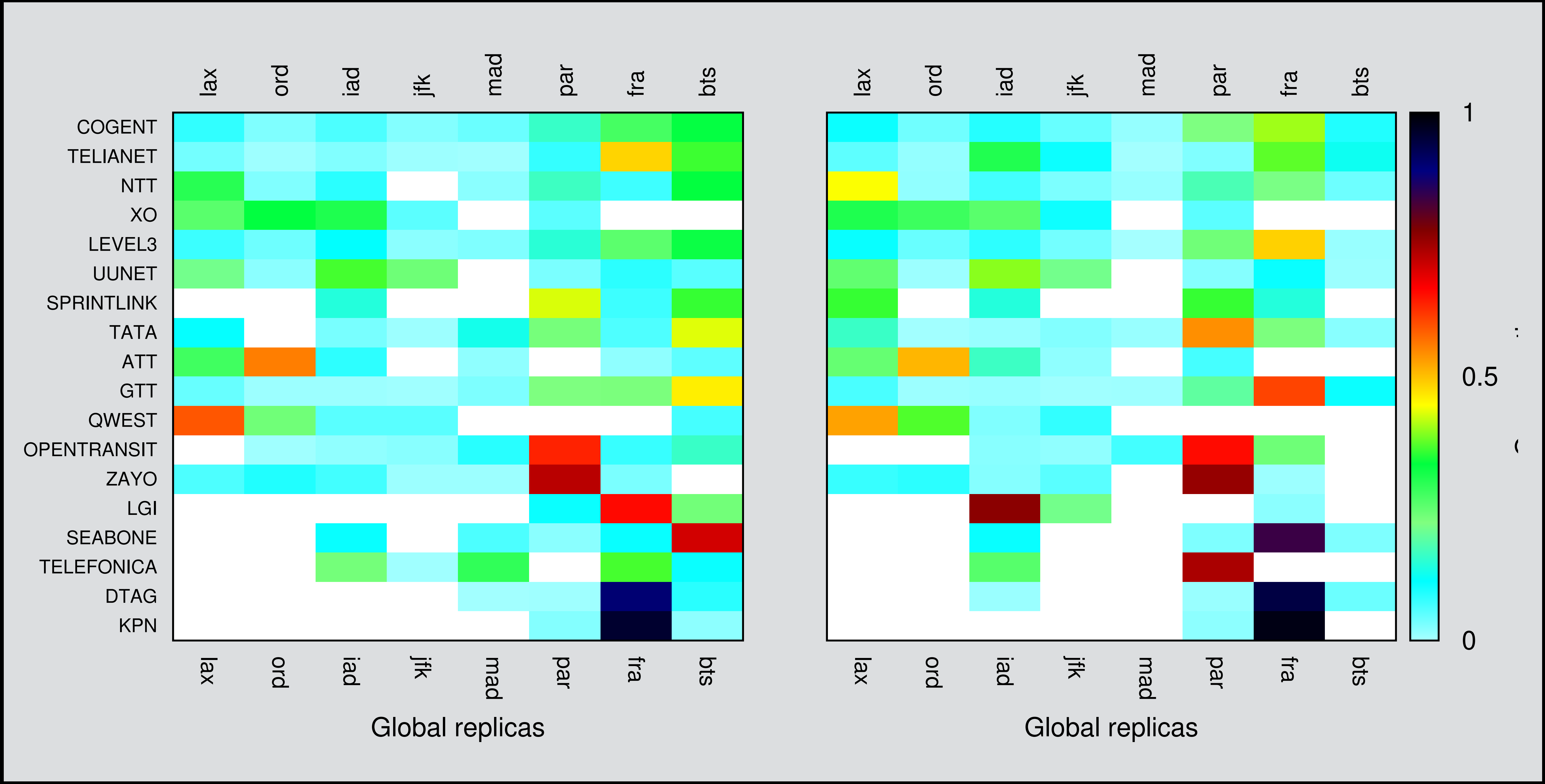
A-Root

- Better. Notably, DTAG sends to London, not Frankfurt.



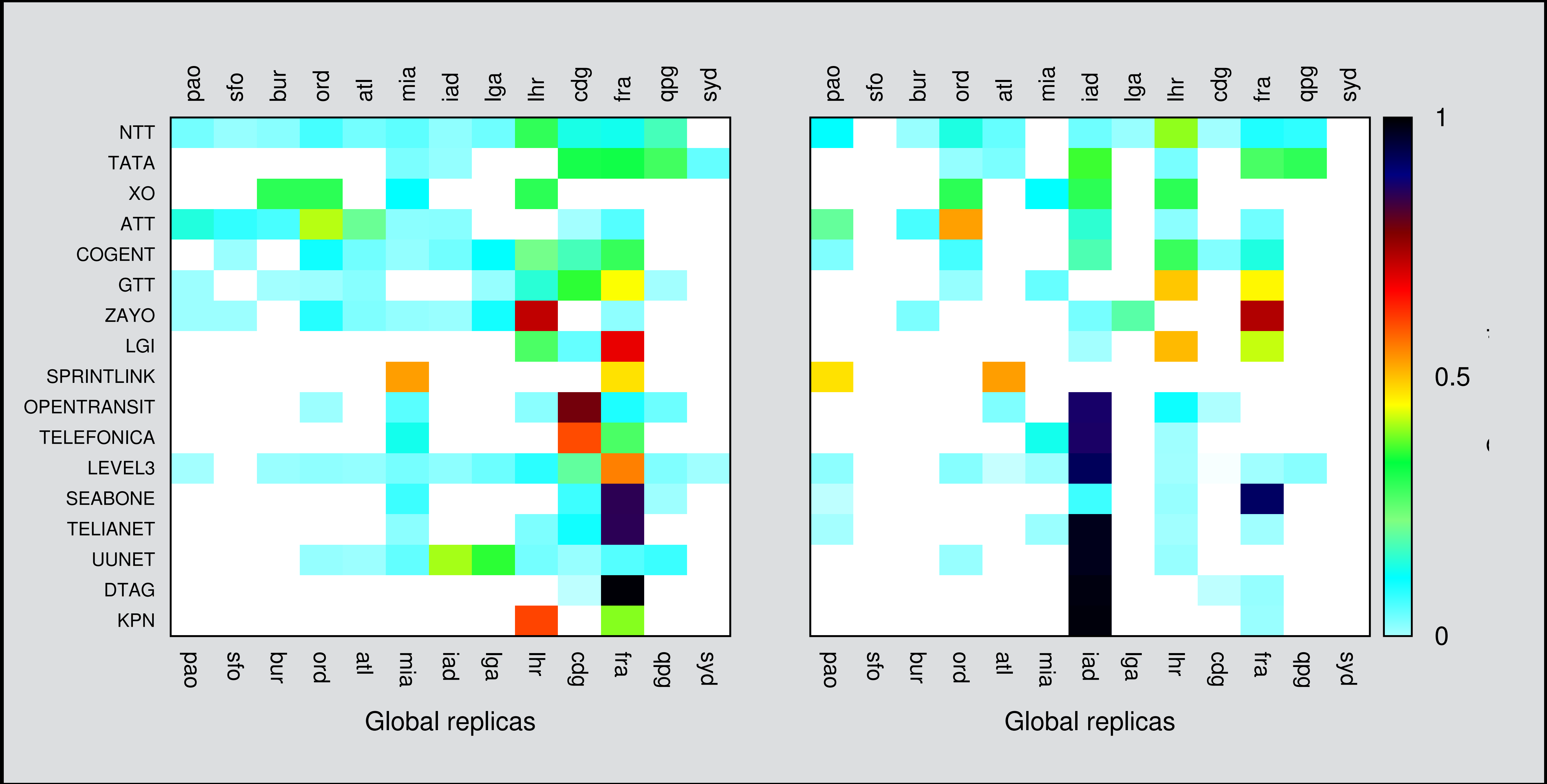
C-Root

- The best at matching tier-1-carried queries to a nearby site.



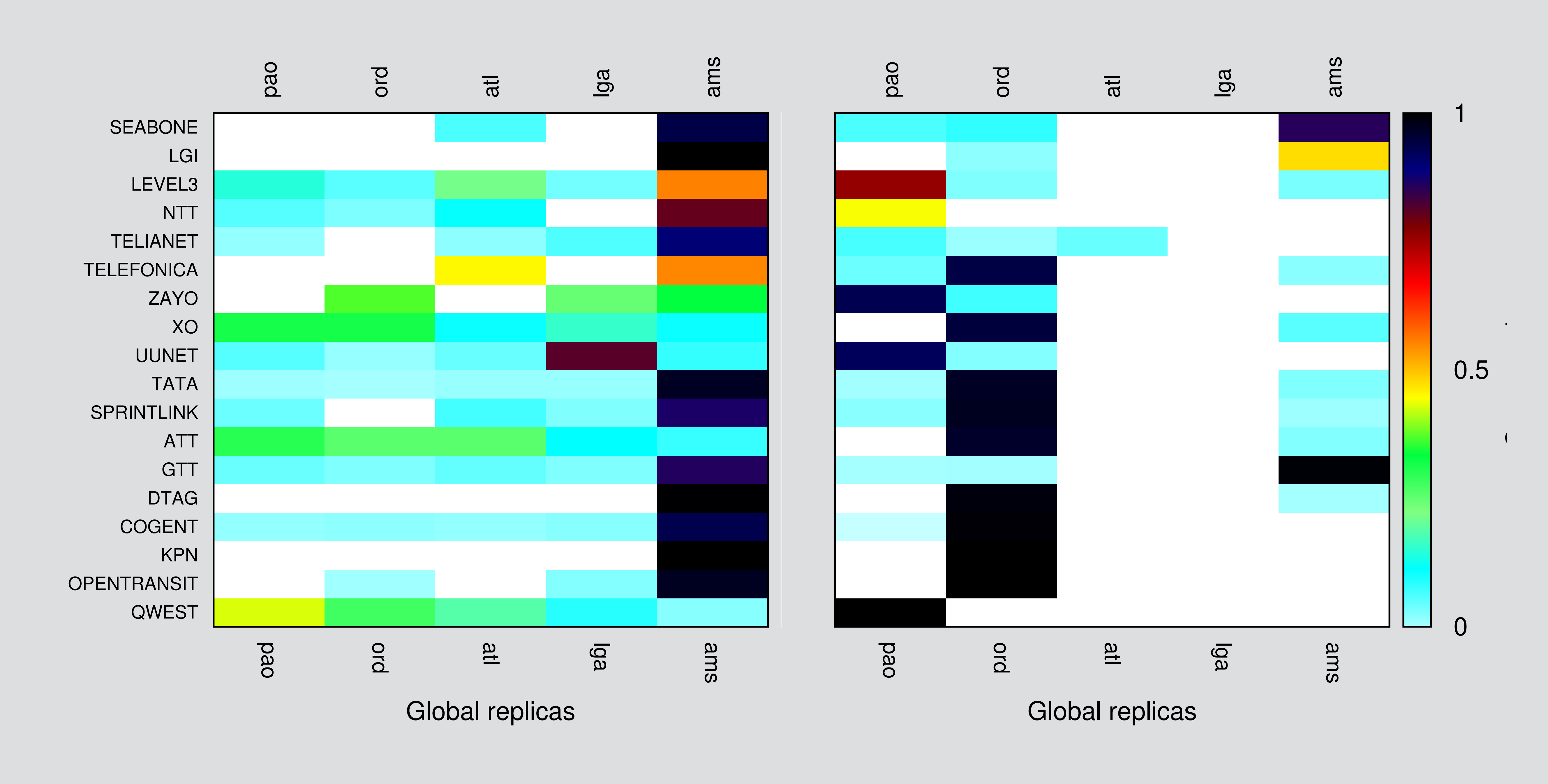
E-Root

- Similar to D in that northern Virginia is preferred, despite Paris, Frankfurt, London query sources.



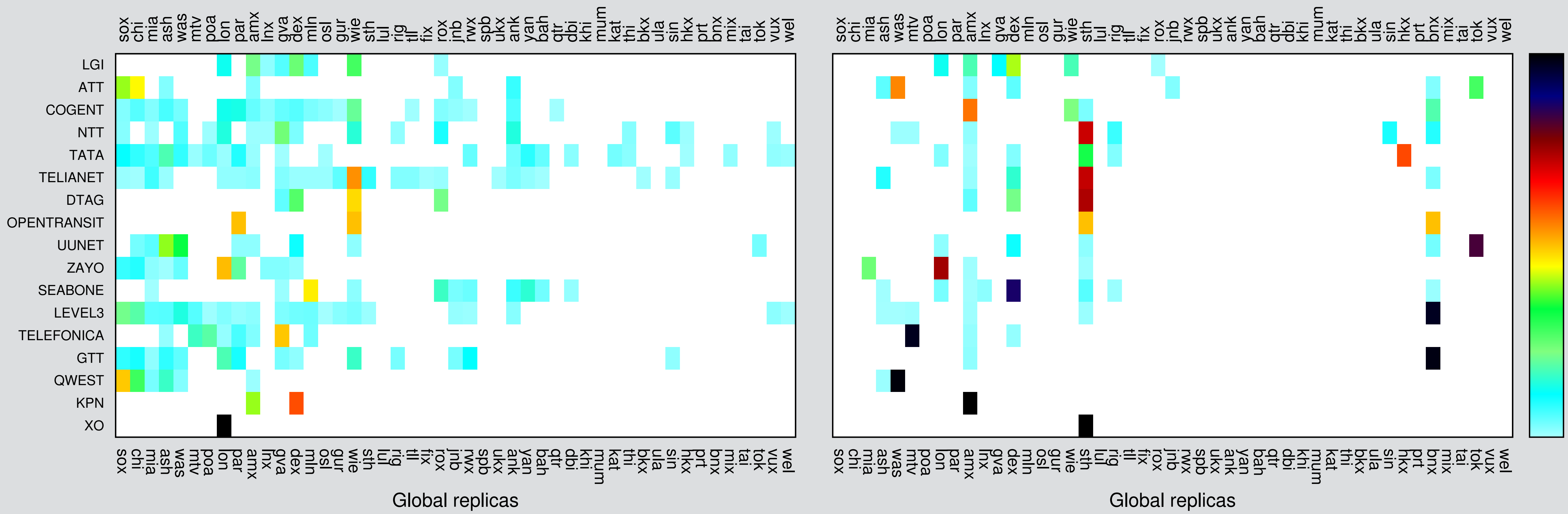
F-Root

- Mostly European RIPE probes served by Chicago despite an Amsterdam replica.



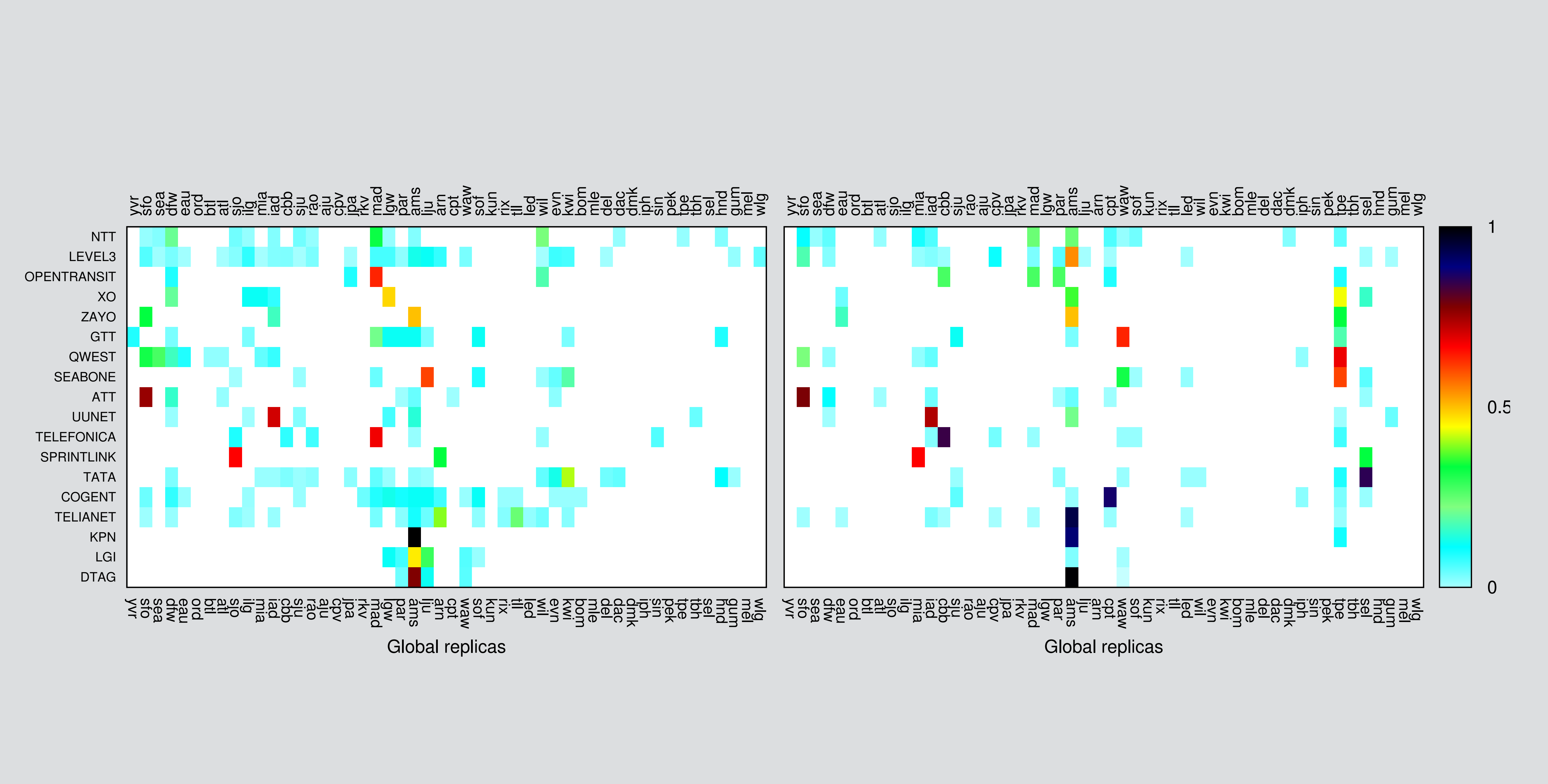
i-Root

- Still picking just one server, not typically the server with the most clients.



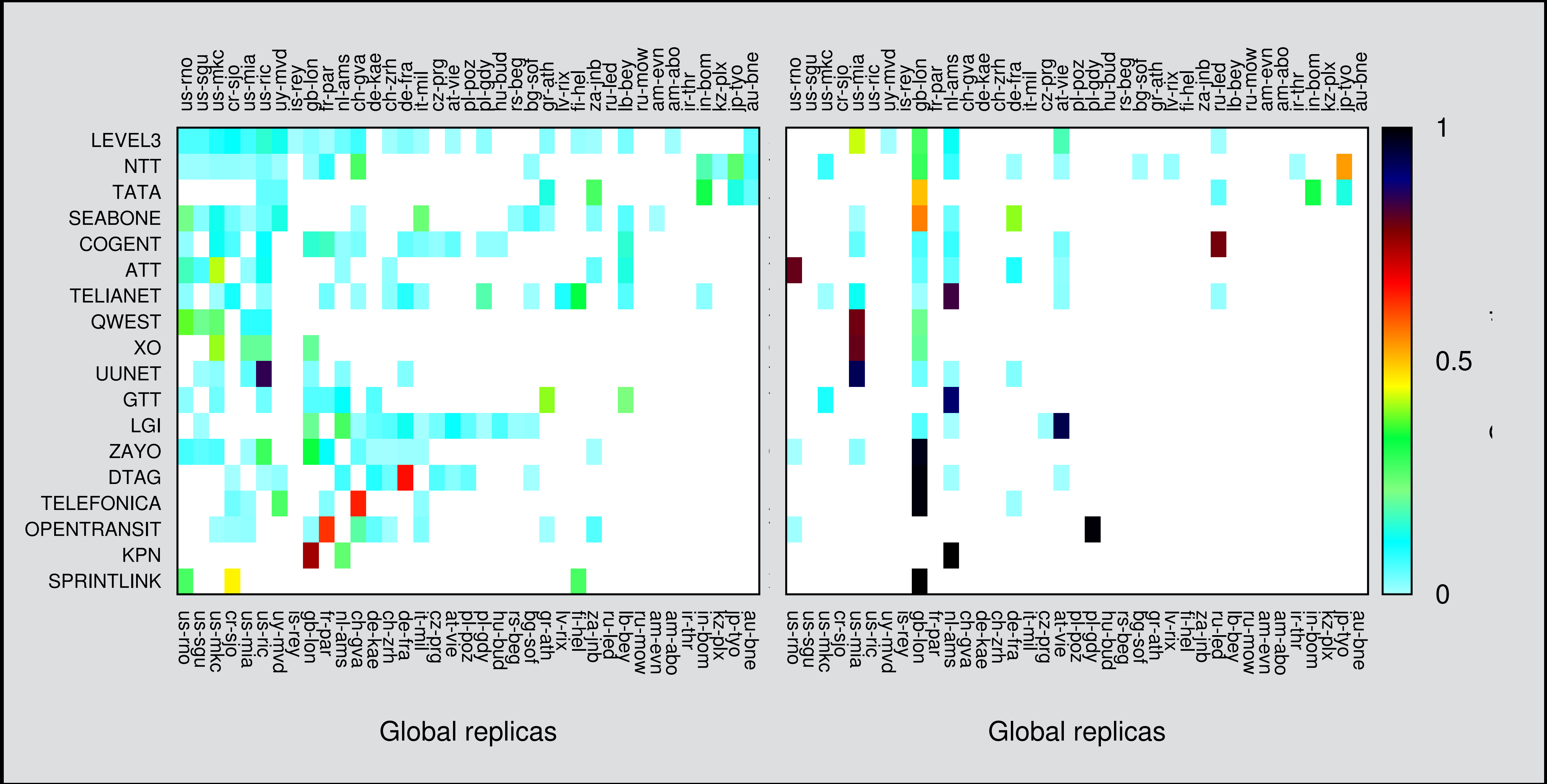
J-Root

- Fairly good, although preference for “tpe” despite no clients.



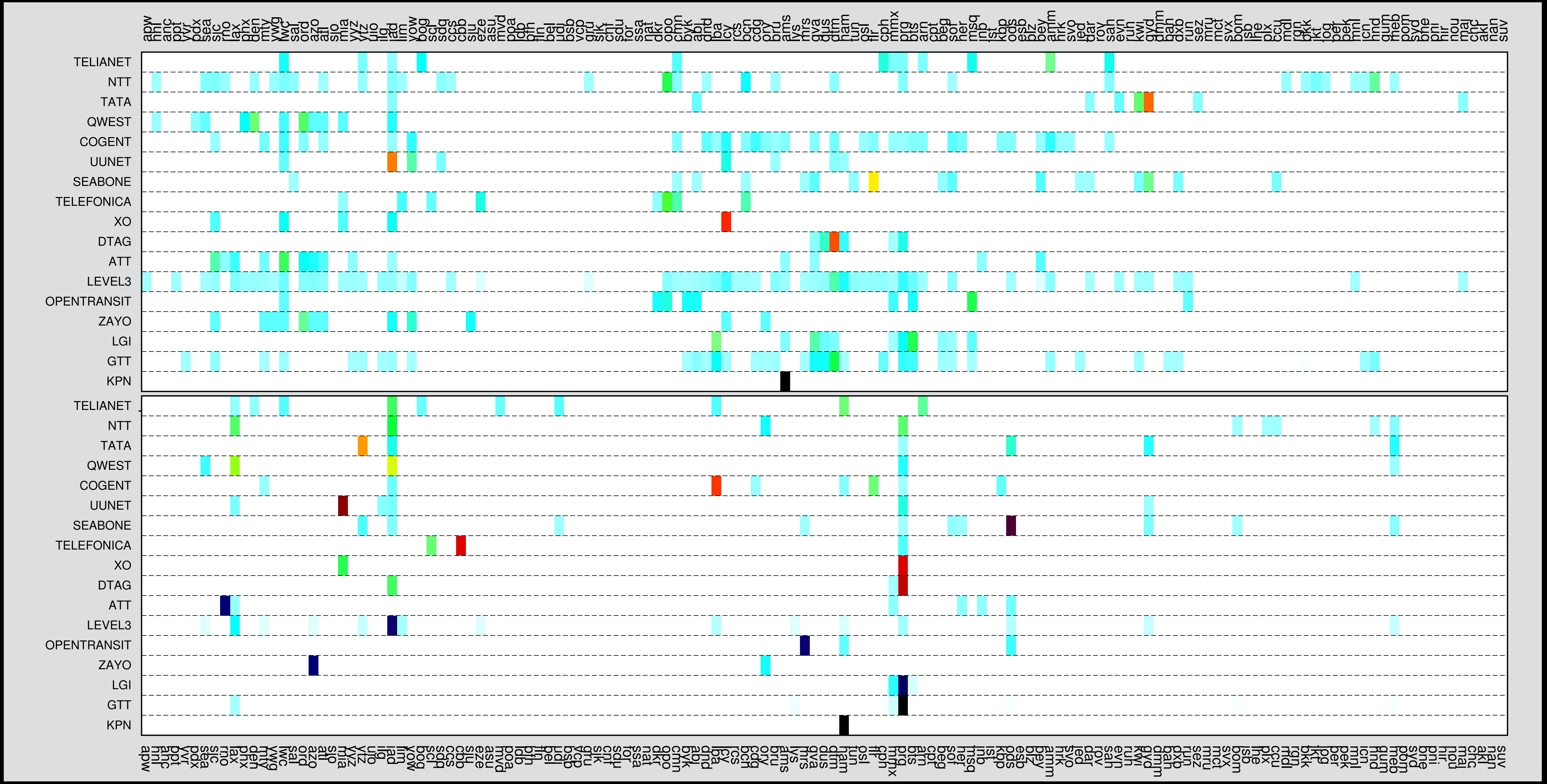
K-Root

- Looks a bit like D.



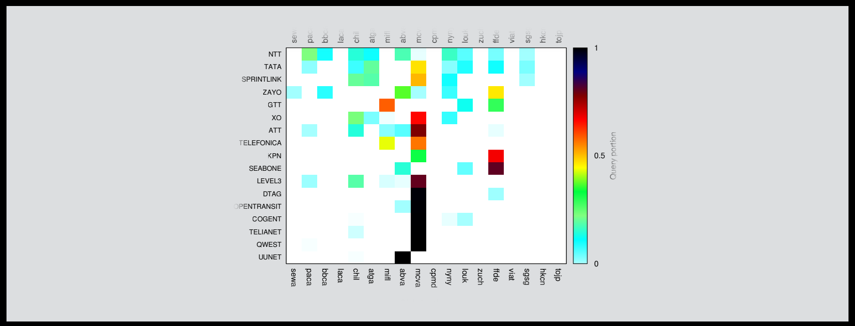
L-Root

- Many global replicas (like i), not often choosing nearby replicas



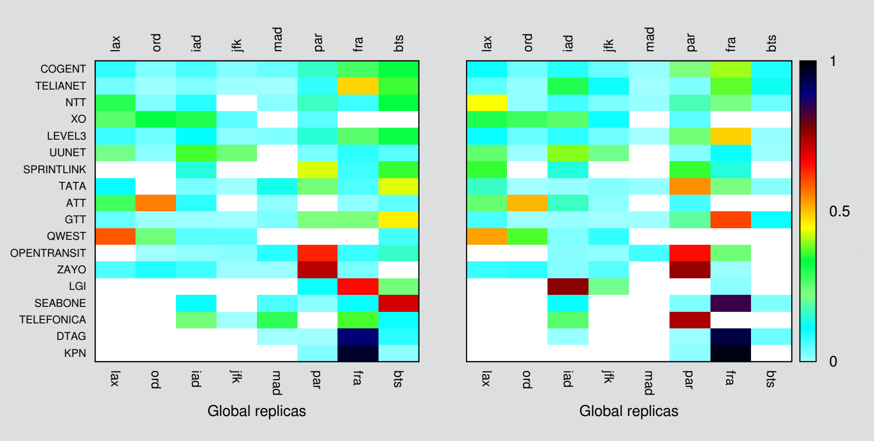
Why is D-Root not distributed?

- ‘mcva’ and ‘cpmd’ are announced through UMD / MAX-Gigapop, which peers with Quest, Telia, Level3. Other replicas are announced by Packet Clearing House (PCH).
- Some Tier-1 ISPs peer only with UMD, thus route queries only to ‘mcva’ and ‘cpmd’.



Why is C-Root so good?

- C is operated by Cogent, another Tier-1
- Expect other tier-1's peer with Cogent widely
- Expect their early-exit-ed queries to go immediately to Cogent, and reach the nearest replica



So how can anycast improve?

(Pretending that my affiliation with Maryland makes me
vaguely responsible for administering this resource)

- Do we bug tier-1 operators?
- Do we assume it’s no big deal since PowerDNS will pick among the 13?
- Do we spend resources elsewhere?